Segmentation of the Accentual Phrase in Seoul Korean

Hae-Sung Jeon, Francis Nolan

Department of Linguistics, University of Cambridge, UK

hsj24@cam.ac.uk, fjn1@cam.ac.uk

Abstract

Pairs of phonemically identical utterances with different location of an Accentual Phrase boundary were presented to listeners. When duration and/or F_0 were swapped between the utterances within a pair, only F_0 change elicited changes in listeners' responses. This effect was found regardless of the distribution of strong consonants which raise Accentual Phrase initial F_0. On the other hand, listeners seemed to be sensitive to a few cases with segmental-prosodic mismatches.

Index Terms: Korean, Accentual Phrase, speech segmentation

1. Introduction

There are multiple acoustic cues available which help listeners in parsing speech [1], and the weight of each cue may differ depending on listeners' linguistic backgrounds (e.g., [2]).

The main goal of this experiment was to investigate the role of F_0 contours and speech timing in parsing word-sized units in Seoul Korean (henceforth Korean) speech.

2. Korean prosody

Korean prosody has been described as similar to that of Parisian French, in that Accentual Phrases (APs) are delimited by intonational contours [3], and the word-internal metrical structure associated with the stress found in Germanic languages is lacking.

Korean is unusual in that the initial tone of an AP tends to be determined by the type of the AP initial segment ([4] and references therein). That is, when an AP begins with a segment with the feature value [+ stiff vocal cords], such as aspirated consonants or fortis consonants (strong consonants hereafter), the AP tends to begin with relatively high F_0, but otherwise it would begin with low F_0. Fourteen types of F_0 pattern are reported [4] as possible within the AP; although F_0 can fall throughout an AP, the rising pattern in an AP seems far more common. In a corpus study [5], 88% of APs which were not in Intonational Phrase (IP) final position had a final rise.

On the other hand, studies in Korean speech timing have reported that there is a substantial final lengthening effect of a large group in speech, such as the IP, while the results regarding the AP have not been consistent (see [6]).

In some recent studies which investigated the role of the F_0 contour and phrase-final lengthening in perception of the AP, the facilitating effect of intonational cues on lexical segmentation is clearly shown [5, 7]. Regarding duration, when the final syllable of the AP preceding the target was lengthened, accuracy in an artificial language learning experiment increased [5]; however, a larger degree of the preboundary lengthening had an effect only when the preboundary tone was a relatively infrequent one (i.e., the infrequent L boundary tone) [7].

In these studies, the possible frequency effect of the intonation patterns (e.g., the commonly occurring H boundary tone) was specifically tested under the assumption that intonation provides a crucial cue for the detection of the word-sized unit in running speech. However, the timing factor was restricted to the phrase-final lengthening, although it is possible that speakers may adjust the duration of non-phrase-final syllables as well, and the experimental materials had only sonorants or lenis stops which induce phrase-initial low F_0.

In the present experiment, the contribution of F_0 vs. timing and how robust their effects are over the different types of segments in resolving segmentation ambiguity are investigated. Given that the factors determining the variability shown both in timing and in the phrase-level F_0 contours of Korean are not fully understood and that research into the perception of Korean APs with strong segments is rare, we began with an exploratory study, manipulating naturally spoken materials using resynthesis rather than synthesising stimuli from scratch.

3. Experiment

3.1. Method

3.1.1. Materials

Set	Group	Phonemic transcription	Phrasing	Meaning
A	1	/imanimano/	2 + 3	20000, 20005
			3 + 2	20002, 10005
	2	/imanipɜko/	2 + 3	20000, 205
			3 + 2	20002, 105
	3	/imankumano/	2 + 3	20000, 90005
			3 + 2	20009, 10005
B	4	/imanitsʰʌno/	2 + 3	20000, 2005
			3 + 2	20002, 1005
	5	/imanosipo/	2 + 3	20000, 55
			3 + 2	20005, 15
	6	/imankutsʰʌno/	2 + 3	20000, 9005
			3 + 2	20009, 1005
C	7	/imansamano/	2 + 3	20000, 40005
			3 + 2	20004, 10005
	8	/imantsʰilmano/	2 + 3	20000, 70005
			3 + 2	20007, 10005
	9	/imansampɜko/	2 + 3	20000, 305
			3 + 2	20003, 105
D	10	/imansamtsʰʌno/	2 + 3	20000, 3005
			3 + 2	20003, 1005
	11	/imantsʰilsipo/	2 + 3	20000, 75
			3 + 2	20007, 15
	12	/imanpʰaltsʰʌno/	2 + 3	20000, 8005
			3 + 2	20008, 1005

Table 1: Test IPs. Commas mark the boundary between two different numbers.

Experimental materials were selected from numbers, since the combination of numbers can form word-like units (see Table 1). There were 12 ambiguous digit strings which could refer to different sequences of numbers depending on the location of a phrase boundary. The word-sized target numbers and the ambiguous digit strings (i.e., the sequences of the two numbers) are referred to as test APs and test IPs respectively in this paper. All test IPs consist of five syllables, which form two APs of either 2 + 3 or 3 + 2.

There were four SETS of test IPs (A – D), classified depending on the distribution of the strong consonants within an IP (see Table 2).

SET A	$\sigma 1$ - $\sigma 2$ - $\sigma 3$ - $\sigma 4$ - $\sigma 5$	none
SET B	$\sigma 1$ - $\sigma 2$ - $\sigma 3$ - $\boldsymbol{\sigma 4}$ - $\sigma 5$	3 + 2
SET C	$\sigma 1$ - $\sigma 2$ - $\boldsymbol{\sigma 3}$ - $\sigma 4$ - $\sigma 5$	2 + 3
SET D	$\sigma 1$ - $\sigma 2$ - $\boldsymbol{\sigma 3}$ - $\boldsymbol{\sigma 4}$ - $\sigma 5$	both

Table 2: Four SETS with different location of the strong segments which raise AP-initial F0. Syllables with strong onsets are marked in bold. The last column shows the case when the second AP is likely to begin with raised F0.

The first author (a native Seoul Korean speaker, female) read all test IPs between the carrier phrases (/ipʌne#tɨlisil#kʌsin#*test IP*#imnita/, meaning 'this time, what you will hear are *test IP*') in four types of phrasings (i.e., 2 + 3, 3 + 2, 1 + 2 + 2, 2 + 2 + 1). Only 2 + 3 and 3 + 2 were used for generating test IPs, and the rest were used to make fillers. The speaker maintained constant voice quality and the IP-final boundary tone (H%) throughout the recording.

All recording was done with a Marantz PM 0670 digital recorder and a Sennheiser MKH 40 cardioid microphone. The sampling rate was 44.1 KHz.

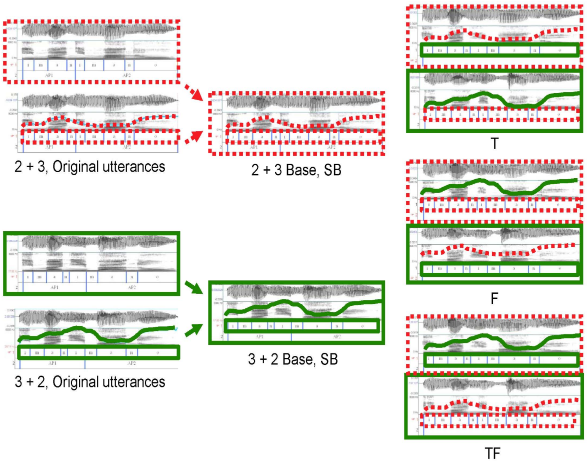


Figure 1: Manipulation of test IPs. The red broken lines represent 2 + 3 phrasing, and the green straight lines do 3 + 2 phrasing of the original utterances. The F0 contour is marked as a line over the spectrogram, and the box in the interval tier around the acoustic intervals represents the temporal structure. There were 8 test IPs (2 SB + 2 T + 2 F + 2 TF) for each GROUP.

From the recorded speech, firstly, two original utterances for each of the 2 + 3 and 3 + 2 phrasing for each ambiguous digit string were selected (Original utterances in Figure 1). Then the duration and F0 of each acoustic interval were replaced by those of another token, i.e., duration and F0 swapped between two utterances with the same phrasing. From the newly resynthesised utterances, one utterance for one phrasing was chosen to have one SB (Synthesised Base) IP. This was done so that all stimuli would be manipulated to have similar sound quality across all of them. In this process, care was taken to have the two SB utterances which were not drastically contrasted in terms of segmental properties. Then the whole F0 contour and/or the temporal structure of the SB IPs were taken and swapped between two IPs with different location of an AP boundary to finally yield 2 test IPs maintaining original prosody (SB), 2 test IPs with timing swapped (T), 2 test IPs with F0 swapped (F), and 2 test IPs with both timing and F0 swapped (TF). Therefore, there were in total 8 test IPs for each GROUP (see Table 1). Note that GROUP refers to a family or some members of a family of resynthesised utterances created based on the same ambiguous digit string. The duration of all test IPs was matched to be 850 ms, and when embedded between carrier phrases, the speaking rate was 6.4 sylls/sec (normal rate, total utterance duration with carrier phrases 2.5 s).

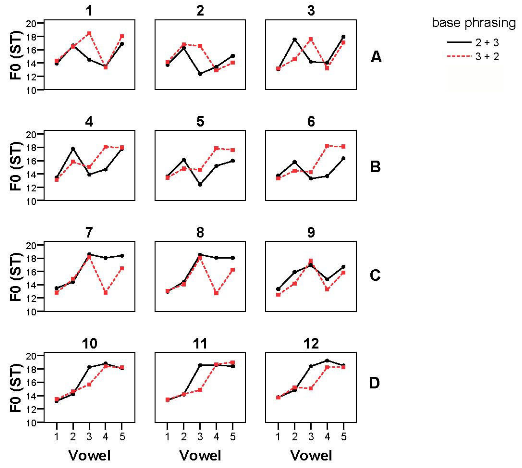


Figure 2: Mean F0 of each vowel in SB IPs by GROUP (semitones). Each row shows GROUPS in SET A (no strong consonant), B (strong onsets in the fourth syllables), C (strong onsets in the third syllables) and D (strong onsets both on the third and the fourth syllables) respectively.

Figure 2 displays the mean F0 of each vowel for all SB IPs. The raising effect of the strong AP initial consonants is observed in the second APs of 3 + 2 SET B, 2 + 3 SET C and both phrasings of SET D. It seems the distribution of the strong segments not only affects the initial F0 of an AP but the overall shape of the F0 within the IPs, as described as a 'see-saw' effect in [8].

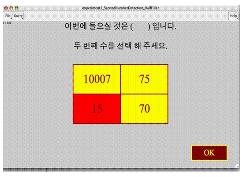
On the other hand, 2 + 3 SB IPs and 3 + 2 SB IPs showed differences in their temporal organisation. The most important finding is that, when duration of the four Inter-Vowel-Onset (IVO) intervals between the five syllables was measured, the duration of the fourth interval was significantly longer in 3 + 2 SB IP ($M = 181.32, SD = 158.02$) than 2 + 3 SB IPs ($M = 158.02, SD = 20.50, p < 0.001$, paired-sample t-test). It may be because the effect of the IP-final temporal deceleration was confounded with the temporal expansion of the second AP for the 3 + 2 SB IPs, due to the smaller number of syllables.

3.1.2. Experimental procedure

When the experiment began, four options appeared on the screen first (see Figure 3), and a sentence was automatically played after 1 second. Test IPs were embedded between carrier phrases (/ipʌne#tɨlisil#kʌsin#*Test IP*#imnita/).

The listeners' task was to identify the second number (i.e., the second AP) from each test IP after listening. In order to keep the real task implicit, listeners were provided with four choices for the second Korean number as determined by the alternative phrasings 1 + 2 + 2, 2 + 3, 3 + 2, and 2 + 2 + 1 (fillers were actually phrased into either 1 + 2 + 2 or 2 + 2 + 1). Since each of the four phrasing options determined a different second number, it was possible to infer from their responses how listeners phrased the test IPs.

An OK button appeared when a listener made a choice for each question. The next trial started only when the OK button



이번에 들으실 것은 (   ) 입니다.

두 번째 수를 선택 해 주세요.

10007	75
15	70

Figure 3: Display on the screen. The carrier phrases and the instruction were written in Korean.

was pressed, allowing participants to change their initial choice.

The same utterance was played for four times throughout the experiment, and the presentation order of all stimuli was randomised for each subject. The same utterance was not allowed to be played twice consecutively.

There was a short practice session before the main experiment, and there were 576 trials (384 trials for the test stimuli, and 192 trials for the fillers) in total. Participants were allowed to take a break after listening to each 48 stimuli if they wanted. On average, it took them 62.55 mins ($SD = 11.82, 1.74$ mins per trial) to finish the experiment and they were given an honorarium for their participation.

Twenty native Seoul Korean speakers participated in the experiment. All experiments were run with Praat.

3.2. Results

3.2.1. Timing vs. F₀ in segmentation of the Accentual Phrase

Factor/interaction	Coefficient	SE	Odds ratio
Intercept	0.46	0.57	1.58
BASE	-3.72***	0.21	0.02
TIMING	0.002	0.06	1.00
F0	-2.38***	0.18	0.09
SET B	0.85***	0.22	2.34
SET C	0.02	0.21	1.02
SET D	-0.63**	0.21	1.88
F0 × SET B	-0.27	0.26	0.77
F0 × SET C	0.53*	0.25	1.70
F0 × SET D	0.37	0.25	1.44
BASE × F0	4.79***	0.27	120.71
BASE × SET B	0.69*	0.28	1.99
BASE × SET C	1.46***	0.28	4.30
BASE × SET D	1.60***	0.27	4.94
BASE × F0 × SET B	0.45	0.37	1.566
BASE × F0 × SET C	-0.89*	0.37	0.41
BASE × F0 × SET D	-0.77*	0.36	0.46

Table 3: Summary of the fixed effects in the generalised linear mixed model (N = 7680, 20 subjects × 4 SETS × 3 GROUPS × 2 BASE × 2 T × 2 F × 4 Repetitions). The reference level was 2 + 3 BASE, SET A, SB. The interactions were not modelled for TIMING. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

For the dependent variable in a dichotomy of 2 + 3 response vs. the rest, a generalised linear mixed model was created with four fixed factors: BASE (i.e., the phrasing of the original utterance, 2 + 3, 3 + 2), TIMING (maintained, swapped), F₀ (maintained, swapped), and SET (A, B, C and D), with all the two- and three- way interactions, and two random factors: SUBJECT, and GROUP (1 – 12) nested in SET. The estimated coefficients and odds ratio of the fixed effects are summarised in Table 3. The odds ratios are the estimation of how likely the response is to be affected by a change of one unit in a certain parameter from the reference level (e.g., when TIMING is swapped). What is important is that most of the main effects and the interactions, except for the effect of TIMING, are significant contributors in this model.

In pairwise comparisons done with data split for each SET, none of the comparisons of SB vs. T and F vs. TF showed statistically significant differences, whereas all SB vs. F, SB vs. TF, T vs.TF pairs were significantly different from each other (all $ps < 0.05$, the Wilcoxon signed-rank test with the Bonferroni corrections). That is, the frequencies of 2 + 3 responses were, when BASE was 2 + 3, SB = T > F = TF, but when BASE was 3 + 2, SB = T < F = TF (see Figure 4). These results confirm the negligible effect of the timing manipulation and that it is the change of the F₀ contour which affected listeners' responses.

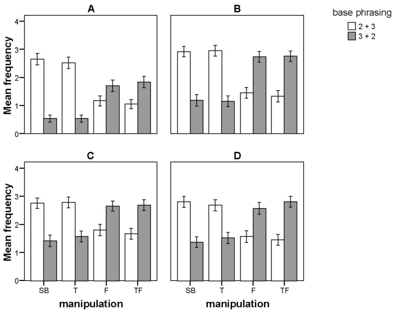


Figure 4: Mean frequency of the 2 + 3 responses for SET A – D. Error bars indicate 1 standard error.

Then, more statistical tests (generalised linear models and the Wilcoxon-signed rank tests) were conducted in order to filter out the GROUPS showing, in one way or another, response patterns different from the rest. The details of the statistical analyses are not discussed further in this paper. To summarise the results, firstly, it was revealed that listeners were strongly biased towards the 2 + 3 response for GROUP 9. The overall F₀ shapes of SB IPs in this GROUP (Figure 2) seem to explain why listeners' responses were peculiar. Despite the different location of the AP boundary, they share nearly identical rising-falling-rising contours.

Secondly, the other slightly deviant one was GROUP 3: there were significantly more 2 + 3 responses for GROUP 3, 3 + 2 F ($M = 2.25, SD = 1.55$) than for GROUP 1 ($M = 1.40, SD = 1.39, N = 20, z = -2.738, p < 0.05$), although no other pairs of GROUPS in each of SB, T, or TF showed significant differences from each other. What is noticeable from Figure 2 is that 2 + 3 SB in GROUP 3 has wider F₀ ranges for both APs (AP1, 4.46 ST; AP2 3.91 ST) than those of GROUP 1 (AP1, 2.73 ST; AP2, 3.37ST). This larger degree of fluctuation of F₀ might have contributed to the better perception of the 2 + 3 phrasing in this case.

3.2.2. (Mis)match of the segmental-prosodic information

We were also interested in whether there were any effects of the mismatch of the segmental and prosodic information. That is, the members of each of the pairs 2 + 3 TF vs. 3 + 2 SB and 3 + 2 TF vs. 2 + 3 SB shared the same F₀ contour and the temporal structure but differed in segmental properties, i.e., in TF, segmental and prosodic information were completely mismatched, whereas in SB, these matched. Therefore, if listeners were sensitive to such mismatch, the response frequencies for SB and TF conditions would be significantly different from each other.

Statistical analysis revealed that there was no statistically significant difference between 2 + 3 TF and 3 + 2 SB. However, when 2 + 3 SB and 3 + 2 TF were compared, there were 4 pairs, GROUP 2, 3, 5, and 7, where the mean frequencies of 2 + 3 responses were statistically significantly higher in SB than TF (see Table 4).

Although the experimenter attempted to select for the resynthesised 2 + 3 SB and 3 + 2 SB utterances which were not drastically different from each other on the segmental level, these four GROUPS may be the ones which the experimenter

Group	2 + 3 SB	3 + 2 TF	z	p
2	3.00 (1.30)	1.70 (1.49)	-3.19	< 0.01
3	2.50 (1.67)	1.90 (1.62)	-2.15	< 0.05
5	3.00 (1.49)	2.60 (1.50)	-1.99	< 0.05
7	2.90 (1.48)	2.40 (1.39)	-2.33	< 0.05

Table 4: Mean (SD) of 2 + 3 response frequencies per subject and the results of the Wilcoxon signed-rank tests of the pairs showing statistically significant differences.

Group	Phonemic transcription	Interval	Difference
2	/imanipɜko/	/pɜ/	more prominent in 3 + 2 TF
			less voicing in /p/, relatively louder /ɜ/
3	/imankumano/	/nk/	[ŋg] in 3 + 2 TF
			[nk^{h}] in 2 + 3 SB
5	/imanosipo/	first /o/	more rounded in 2 + 3 SB
			roundedness dispersed throughout the AP in 2 + 3 SB
		/s/	[ʃ] in 2 + 3 SB
		/si/	relatively louder in 3 + 2 TF
		final /o/	more rounded in 2 + 3 SB
7	/imansamano/	first /a/	more prominent in 2 + 3 SB
		/s/	less voicing in 3 + 2 TF

Table 5: Possible causes of the significant differences in listeners' responses between 2 + 3 SB and 3 + 2 TF. Note that all IPs presented in this Table had the F_0 contour and temporal structure of 2 + 3 phrasing.

failed to achieve sufficient control over segmental cues. In any case, listeners were sensitive to segmental properties in these Groups and we examined some possible cues left in them, as briefly summarised in Table 5.

These segmental cues may possibly be related to the systematic variations of segmental strength depending on their positions within prosodic phrases (e.g., [9]). Although intonational cues were never overridden by segmental information in this experiment, it seems that perceptually salient syllables, such as the ones with less voicing or higher relative intensity, appearing phrase-medially, conflicted with the intonational cue, particularly for Group 2, 5 and 7.

In addition, coarticulatory information seemed to have influence on listeners' judgements: assimilation occurring between the second and the third syllables ([ŋg]) in Group 3 seemed to hinder the choice of 2 + 3. Interestingly, the presence of a prosodic boundary seemed to inhibit the roundedness of a vowel spreading throughout the AP in Group 5, allowing the segments within the AP to be more coherent, and this might have been another cue for AP detection.

4. Discussion

The present experiment compared the contribution of speech timing and the F_0 contour to the perceptual segmentation of utterances into APs, by resynthesising stimuli with the properties of utterances produced with two alternative parses. The result is that the effect of timing was insignificant, and it was the F_0 contour which cued the grouping of word-sized units in speech. It is notable that the strong F_0 effect was consistently found across different segmental distributions.

Some possible causes of such little effect of timing are, firstly, that the durational difference in IVO intervals between 2 + 3 SB IPs and 3 + 2 SB IPs did not exceed 30ms on average, and such a minor difference in duration might not affect listeners' perception of the utterance's overall temporal structure. Secondly, it should be noted that listeners were not asked to make quick phrasing decisions. It is likely that listeners tried to memorise the stimuli's acoustic information before making the final decision, and speech melody may be a feature which is more effectively stored in their memory and retrieved, compared to speech timing information. Thirdly, Korean speakers may be more flexible with speech timing on the level of the word-sized units compared to speakers of other languages, as reflected in the ongoing controversy and inconsistent findings on rhythmic timing characteristics in Korean speech (e.g.,[10, 11]). It is possible that timing or duration is used for some other purposes, such as expressing stylistic differences or emphasised elements in speech, rather than as a primary cue for parsing word-sized units.

In addition, although possible cues for AP segmentation on the segmental level were not originally intended to be included in this experiment, perceptual salience of a certain syllable or coarticulatory information which was not controlled appeared to affect listeners' judgements in a few cases. These cues did not seem strong enough to override the F_0 cue; however, not only their possibly independent roles in speech segmentation but how they conspire with F_0 cues seem to deserve further investigation.

5. Conclusions

In Korean, F_0 plays the primary grouping function not only in production but also in perception of speech. Despite the systematic difference of the temporal structure between 2 + 3 and 3 + 2 phrasing, listeners reacted little to the change in timing. On the other hand, there were cases where information on the segmental level seemed to influence listeners' judgements.

6. Acknowledgements

We would like to thank Geoffrey Potter for technical assistance in the Phonetics Laboratory, Cambridge, and the members of the UCL Speech Laboratory for letting us use their facilities and for their help.

7. References

[1] Christophe, A., Gout, A., Peperkamp, S., and Morgan, J., "Discovering words in the continuous speech stream: the role of prosody", Journal of Phonetics, 31:585–598, 2003.

[2] Cumming, R., "A two-dimensional, cross-linguistic study of speech rhythm", PhD thesis, University of Cambridge, in progress.

[3] Jun, S.-A. and Fougeron, C., "A Phonological Model of French Intonation", in Intonation: Analysis, Modeling and Technology, Botinis, A. (ed.) Kluwer Academic Publishers, 209–242, 2000.

[4] Jun, S.-A. "K-ToBI (Korean ToBI) labelling conventions (Ver. 3.1)", UCLA Working Papers in Phonetics, 99:149–173, 2000.

[5] Kim, S., "The Role of Prosodic Phrasing in Korean Word Segmentation", PhD thesis, UCLA, 2004.

[6] Chung, H., "Analysis of the Timing of Spoken Korean with Application to Speech Synthesis", PhD thesis, University College London, 2002.

[7] Kim, S. and Cho, T., "The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean", JASA, 125:3373–3386, 2009.

[8] Jun, S.-A., "Influence of micorprosody on macroprosody: a case of phrase initial strengthening", UCLA Working Papers in Phonetics, 92:97-116, 1996.

[9] Lee, E.-K., "Acoustic effects of prosodic phrasing on domain-initial vowels in Korean", in Proceedings of the 14th ICPhS, Saarbrücken, Germany, 1141–1144, 2007.

[10] Ross, T. and Arvaniti, A., "Rhythm metrics explain little about speech timing and rhythm", manuscript.

[11] Yun, I., "A Study of Timing in Korean Speech", PhD Thesis, University of Reading, 1998.